IN THE SPECIFICATION:

Please revise the specification as follows:

Revise the first full paragraph on page 2, as follows:

One potential benefit to be gained ~~form~~ from helpdesk data sets is to "mine" them to discover general categories of problems. Once a meaningful "problem categorization" has been discovered, individual categories can be studied to find automated solutions to address future user problems in this area. A typical example of such a solution would be an entry in a "Frequently Asked Questions" section of a customer support web site.

Revise the second full paragraph on page 3, as follows:

Figure 4 shows the floating point format of the same dense matrix (e.g., the preferred format for data mining algorithms). This floating point format represents the same information as the integer format but each document is now “normalized” into unit vectors, thereby eliminating the effect of document length. As can be easily seen, each floating point number is the integer value multiplied by the reciprocal of the square root of the summation of integer values squared, the well known process of normalizing a vector, where the vector is considered as a matrix row. If the document corpus contains a large number of short documents and ~~that~~ the dictionary contains a large number of terms, then it is easy to see that the dense matrix representation would be filled mostly with zeroes since any row representing a document would contain only a few of the many dictionary terms. The matrix in Figure 4 takes 48 bytes in RAM, assuming a short integer takes two bytes and a floating point number takes four bytes.